DOCUMENT RESUME

ED 035 322 FL 000 757

AUTHOR	Pimsleur, Paul
TITLE	A French Speaking Proficiency Test.
INSTITUTION	American Association of Teachers of French.
PUB DATE	Apr 61
NOTE	10p.; Reprint from The French Review, v34 n5 p470-479 Apr 1961
EDRS PRICE	EDRS Price MF-$0.25 HC-$0.60
DESCRIPTORS	*Audiolingual Skills, *French, Language Fluency, *Language Proficiency, *Language Tests, Pictorial Stimuli, Pronunciation, Speaking, Syntax, Tape Recordings, *Test Construction, Test Reliability, Test Validity, Vocabulary

ABSTRACT

An attempt to test students objectively in a five-part, French, speaking proficiency test is described and discussed. Concrete nouns, abstract words, pronunciation, syntax, and fluency are tested with a combination of tape and picture stimuli. Reliability, validity, and practical questions are raised; and previous aural-oral testing procedures are reviewed. (AF)

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
OFFICE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY.

Reprinted from THE FRENCH REVIEW
Vol. XXXIV, No. 5, April, 1961

ED035322

A French Speaking Proficiency Test

by Paul Pimsleur

A. *Introduction*

At the present juncture in FL teaching it is hardly necessary to point out the need for tests of speaking proficiency. Much depends on such tests. The difference between merely paying lip service to the oral goal and actually achieving it resides in making clear to the students that their marks will depend in considerable measure upon their performance in speaking the foreign language. This performance may be checked at various intervals and in various ways, ranging from an informal teacher quiz to a standardized test. This article reports on a standardized test which attemps to go as far as possible in the direction of complete objectivity.

The effort to test speaking ability entirely objectively is doomed to fall short of complete success, for an evaluation of how well a person speaks French requires judgments on the part of a hearer. These are necessarily subjective. Yet the attempt to structure these judgments so as to make them similar from judge to judge will be useful and instructive; it should give us not only a new test, but a clearer idea of what we mean we say someone can "speak French."

It is a simple matter to review past efforts in this area. As of 1953, Furness, reviewing aural-oral tests in Spanish (*MLJ*), reports that no test of either aural or oral proficiency had given serious evidence of validity and reliability. The same can be said for French. This is not to say that there were not worthy attempts made, but the author of a test is obliged to present evidence on validity and reliability in order to be convincing, and this had not been done in the case of any audio-lingual tests. In aural testing this situation has been altered somewhat in recent years (e.g. Brooks, *French Listening Comprehension Test*, Carroll-Ho Chi-Min-Pimsleur, *Pictorial Auditory Comprehension Test*). As for oral testing, little has been done. Kaulfers reported an attempt at objective testing of oral proficiency as long ago as 1944 (*MLJ*). However, his test bore on the general capacity to speak the French language, rather than on the semester by semester learning which takes place in the schools. Consequently it is more valid for measuring, for example,

FL 000 757

a diplomat's ability to converse in French than as a measure of whether or not students have mastered certain elements of the language at the end of a particular semester. The attempt by Stabb (*MLJ,* 1955) deals with school achievement, but the scoring involves sufficient subjectivity so that the test cannot be widely used and scored by untrained judges, nor can it be administered en masse to large groups of students in a limited time. No test now exists which meets the criteria of validity, reliability, ease of administration, and objectivity of scoring.

B. *The French Speaking Proficiency Test*

Now for the *French Speaking Proficiency Test* itself. For purposes of testing, the matter to be tested was broken down into five parts, which are intended to represent the most important aspects of a student's ability to speak French within the limitations of the content of his school course. In order to express himself in French, a student must 1) know the names of concrete objects (e.g. chair, table), 2) know words for abstractions (e.g. empty, night, happy), 3) have a reasonably good pronunciation, 4) command a certain number of syntactic patterns, and 5) feel free to utter his thoughts in French with some ease. These are the 5 parts of the test.

Part One: Concrete nouns. The student sees in his test booklet a number of pictures of things which he must name in French, saying his answers into the microphone to record them. His score is the number of items he can name correctly in 45 seconds.

This part (and Part II which is similar) differs from a conventional vocabulary test, first of all in that the student must recall and say the names of things almost instantly as he is under time pressure, secondly in that the stimulus is a picture rather than an English word, and thirdly in that he has a problem of pronunciation but not of orthography. To avoid confusing the task, the gender of the word need not be said correctly (or at all, for that matter) and the student is so informed in the instructions. The score is the number right. The maximum number of items on part I is 28.

Part Two: Abstract words. The student sees, in his test booklet, a pair of pictures. The first shows a smiling boy, and the caption says "Le garçon est heureux." The second shows the same boy crying, and the caption says, "Le garçon est ————————." The student must say (into the microphone) a word which correctly completes the caption. In the example, the word might be *malheureux, triste, désolé,* etc. By means of such pairs of pictures, abstract oppositions are elicited like

full-empty, night-day, more-less, good-bad. A time limit of one minute is allowed for this part. The score is the number right. The maximum number of items in this part is 16.

While testing the student's knowledge of abstract words, this part also tests his IQ, since he must figure out each item before he can answer it. Part II can no doubt be eliminated in future editions. The resulting reduction in scoring time from five minutes to four minutes per student is a substantial advantage.

Part Three: Pronunciation. The student finds in his test booklet a list of twenty sentences. He is given time to practice them, and then records his reading of them.

1. Il est fou.
2. Il est beau.
3. Nous sommes dans la salle.
4. J'ai vu le bébé cet été.
5. Qu'est-ce qu'il a bu?
6. Regardez le feu.
7. J'en ai neuf.
8. Il me le dit.
9. Ce train est lent.
10. Qu'est-ce qu'ils font?
11. Servez le pain.
12. Paris est grand.
13. Il est à la maison de son oncle.
14. Quelle jolie harmonie!
15. Où est Jean? / Où est Jeanne?
16. Le vin est bon. / La viande est bonne.
17. Mon frère est marin. / Il est dans la marine.
18. J'ai vu la fille. / Elle est en ville.
19. Quel pays! / Il y a du soleil!
20. C'est un jeu. / Je joue.

These twenty items represent some of the important elements of French pronunciation. In the interests of objective scoring, each sentence contains only one element which the scorer must listen to and judge. The first twelve items contain twelve different vowel sounds. The sound in question is usually in the last syllable so that it will receive the tonic accent. Item thirteen tests liaison (Il est‿à la maison de son‿oncle.) Item fourteen tests the silent *h*. Items fifteen through twenty contain oppositions (Jean/Jeanne, bon/bonne, marin/marine,

fille/ville, pays/soleil, jeu/joue) which are among the more difficult ones for American students to maintain.

Here, the scoring becomes a problem, for the scorer must judge the adequacy of the student's pronunciation in each item. Subjectivity necessarily enters, and a scoring system must be found to keep it to a minimum. After examining other possible measuring scales, from a 2-point scale (right or wrong) to a five-point scale (poor, fair, good, very good, excellent), it was decided that a three-point scale would be the best compromise. The scale was numbered 0, 1, and 2, and a description given each score:

2 = like a native
1 = not native but adequate
0 = inadequate

This scale does not require very fine judgments on the part of the scorer, while still permiting a sufficient range of scores. The scorers are instructed to practice on at least ten recordings, so as to stabilize their judgments, before beginning actual scoring. This simple 0, 1, 2, system is also used in parts IV and V of the test. Some evidence concerning the degree of agreement among judges using this system will be presented in section C of this article.

The original version of Part III is the one described here. Subsequent analyses of the individual items have permitted improvement of the test by the elimination of items on which most students get perfect scores (e.g. item 14), and the addition of items containing sounds of proven difficulty. The sentences have also been revised so as to contain *two* examples of the sound in question, one in tonic and the other in non-tonic position.

The first three parts of the test used the test booklet. The remaining two parts are conducted between tape and student, with no printed material.

Part Four: Syntax. The most difficult test to construct was for syntax. It was hoped a picture device could be used, in which a picture of an action would elicit a sentence describing that action. It would be simple to draw reaction-producing pictures, like that of a boy washing his hands. But they would elicit a variety of reactions which would make objective scoring difficult. Hence a more direct approach was utilized.

The students hear sentences in English. He is required to convey each sentence at once in French. The word convey, rather than trans-

late, is used advisedly, for the sentences were selected in such a way that they merely provide an input of information, which information must then be transmitted in French. It is not a translation which is called for, for no word-for-word translation will do the job.

In order to familiarize the student with this technique, he is given some practice in it. Then he hears the following test sentences, to be conveyed in French immediately. Note that each item involves particular syntactic problems, and that these problems are roughly graded in order of difficulty. The vocabulary intentionally presents little difficulty, so as to focus attention on the syntactic problems.

1. Roger has friends.
   Roger a des amis. (avoir; partitive)
2. He doesn't like his friends.
   Il n'aime pas ses amis. (regular verb; possessive)
3. Louise is Roger's little sister.
   Louise est la petite sœur de Roger. (word order; adjective)
4. They go to the same school.
   Ils vont à la même école. (irregular verb word order)
5. He gives her a few books.
   Il lui donne quelques livres. (indirect object; *quelques*)
6. They saw three friends yesterday.
   Ils ont vu trois amis hier. (present perfect)
7. The friends said something to them.
   Les amis leur ont dit quelque chose. (object; tense)
8. But Roger hadn't seen them.
   Mais Roger ne les avait pas vu. (object; pluperfect)
9. They won't speak to him tomorrow.
   Ils ne lui parleront pas demain. (future; negative)
10. Would you like to know Roger?
    Aimeriez-vous connaître Roger? (conditional; interrogative)

This part is scored on the 2, 1, 0 scale, where:

2 = completely correct
1 = partially correct
0 = incorrect or missing

Part Five: Fluency. This part is designed to test the student's readiness to give forth a response in French in a conversational situation. To accomplish this, a simulated situation is created. The student is informed in advance what the conversation will be like. He is told:

"We are going to hold a simple, everyday kind of conversation. I want you to imagine that we are both American students who have gone to Paris. We meet there, quite by accident, on the street. We say hello, then I ask you when you got to Paris, and you answer. I ask you where you live, and you tell me you live with a French family, or in a hotel. Then I ask you what you're doing this evening, and you say you're going to the theatre. I ask you what time the theatre begins, and if you can have dinner with me before going there. You accept, and we agree to meet at the "chez Maxime" restaurant at six o'clock."

The student then records this conversation in French, with the tape taking one role and he himself the other.

Scoring is by the 2, 1, 0 scale, where:

2 = responded promptly and well.
1 = responded promptly but poorly, or, responded hesitantly but well.
0 = responded hesitantly and poorly, or, no response.

C. *Reliability*

The sort of reliability with which we are most concerned asks the question: do different scorers arrive at the same score for a given student? If not, by how much do their results differ? The issue here is one of inter-judge reliability.

Two different kinds of evidence are available which bear on inter-judge reliability. On the one hand, there are results of a number of different scorers doing the same few cases. On the other, we have evidence of two scorers doing many cases. The evidence will be presented in that order.

Three cases were selected at random from among several hundred test recordings. Each of these three was corrected by five different judges —not always the same five in each case (in all, seven different judges are represented). All judges were native speakers of French, who were given a ten-minute training period in how to score the test. The results are presented in Table I. The mean total score was calculated for each subject, and the deviation of each of the five cases from this mean. These deviations were then made into a single distribution, whose standard deviation was calculated. (The correlational factor introduced by the fact that some of the judges made more than one judgment was simply ignored.) In this way, an estimate of the standard error of a test score was arrived at, which turned out to be 3.5.

Table 1

A: Scores given to Student A by 5 different judges

	Part I	Part II	Part III	Part IV	Part V	Total Score
Judge 1	16	8	32	8	11	75
2	17	9	34	9	12	81
3	17	9	35	10	13	83
4	20	7	38	12	11	89
5	17	8	36	8	13	82
					Mean	82.0

B: Scores given to Student B by 5 different judges

	Part I	Part II	Part III	Part IV	Part V	Total Score
Judge 1	15	8	26	8	10	67
2	15	10	27	10	11	73
3	14	8	27	11	10	70
4	13	8	30	11	10	72
6	15	8	27	9	6	65
					Mean	69.4

C: Scores given to Student C by 5 different judges

	Part I	Part II	Part III	Part IV	Part V	Total Score
Judge 1	15	7	22	10	2	56
2	15	7	28	7	6	63
3	14	7	23	8	5	57
6	15	7	23	7	5	57
7	15	7	24	9	6	61
					Mean	58.8

S.E. = 3.48

The standard error of measurement is an estimate of the limits within which we can have confidence that the true score lies. In the case of a familiar score, like an IQ, we know that an IQ of 122 does not mean the person's IQ is *exactly* 122, but rather that it is somewhere in that neighborhood. When informed that the standard error of an IQ is

five points, then we can say with some confidence (about 68% chance of being right) that the person's IQ lies between 117 and 127 (plus or minus one standard error). We can say with even greater confidence (95%) that his true IQ lies between 112 and 132 (plus or minus 2 standard errors).

In the case of our test, a standard error of 3.5 means that a student's score of, let us say, 84 should be regarded as lying between 80.5 and 87.5 (plus or minus 1 S.E.) or, for greater assurance, between 77 and 91 (plus or minus 2 S.E.'s).

The size of standard error of meaurement, 3.5, is highly satisfying. It compares favorable with many widely used tests, and is particularly impressive when measuring an ability so difficult to judge objectively as the ability to speak French. It gives us confidence in the extent of interjudge agreement.

Further confidence may be gained from a different set of data. The test was administered to 34 students whose recordings were then corrected by two different judges. After an initial session in which they agreed on scoring procedure, the two judges independently corrected the 34 recordings. Their two sets of scores correlated to the extent of .93. This is a remarkably high correlation. Looking at these data in another way, the students were placed in rank order, from the highest to the lowest score, as assigned by each of the judges. The rank-order correlation was .91. Granting that not all judges will agree this well, these high correlations still show that the goal of objective scoring has largely been achieved.

It was not possible to obtain information on the kind of reliability called *stability* (the extent to which the test is measuring accurately). Split-half reliability is inappropriate for a speeded test, and alternate forms were not available, nor was it feasible to test and retest the subjects. Such data will be reported when alternate forms become available.

Reliability concerns the accuracy of the test as a measuring instrument. We now turn to the question of validity, which concerns whether the test really measures the thing it claims to measure.

D. *Validity*

There are many kinds of validity. That is, there are many ways of asking whether this test really measures French speaking proficiency. One may merely inspect it to satisfy oneself that the tasks and items have apparently been well chosen (face validity). One may compare results on this test with results on other tests of the same ability (con-

gruent validity). One may examine the test's success in predicting how well a person can speak, as measured by some outside criterion such as the opinion of a French teacher (predictive validity).

This test rests largely on face validity, and can fairly do so because it does not claim to measure anything more than what is measured in the subsections of the test. It does not attempt, like many psychological tests, to infer something about a person's inner workings. It merely structures some French-speaking tasks as a measure of the ability to speak French. The user must decide for himself to what extent he agrees that the items and tasks contained in the test really are relevant to French speaking proficiency.

A point of interest arises in regard to the weighting of the five subparts. Their present weighting is as follows:

Parts I & II (vocabulary) have a maximum of 44 out of 116 points	= 38%
Part III (pronunciation) has a maximum of 40 out of 116 points	= 35%
Part IV (syntax) has a maximum of 20 out of 116 points	= 17%
Part V (fluency) has a maximum of 12 out of 116 points	= 10%
Total:	100%

There might be some debate as to whether vocabulary and pronunciation should count for so much, and whether syntax is not somewhat undervalued. In order to speak French, what is the relative importance of these factors? At the "pidgin" level, vocabulary is more important than either pronunciation or syntax, it would seem. As the level rises, the latter aspects, particularly syntax, become more and more important. Until linguists have clarified this issue, the test maker must use his best judgment in assigning weights to the various factors.

Further evidence on the validity of the test may be cited. In a sample of 33 UCLA students, there was a correlation of .60 between their scores on the FSPT and the grades assigned for their oral work in the language laboratory, the latter being based on many observations during the semester. The two sets of scores are independent, since different scorers are involved. This correlation is satisfying, particularly in view of the unreliability to which teacher grades (in this case, grades assigned by a lab instructor) are subject. If allowance is made for this by assuming a reliability of .80 for the lab grades and of .90 for the FSPT, then the correlation between the two, corrected for attenuation, rises to .71.

E. *Practical Considerations*

The *French Speaking Proficiency Test* takes about 20 minutes to administer. It can be administered to individuals or to groups, but requires certain equipment on the part of the school. It must be played on a tape recorder which feeds into the earphones of each examinee. Every examinee must be seated before his own recording machine. Hence, the size of the group which can take the test at one time is limited by the school's laboratory setup.

The test yields a four-minute recording for each student. These recordings must be corrected, using the scoring sheet. With a little practice, the scorer can judge the recording as it is playing, without having to stop it or repeat.

The test is being revised on the basis of past experience. Present plans call for Elementary forms A and B, suitable for use at the end of one semester in college, or one year in high school, and Intermediate forms C and D, for use thereafter.

In conclusion, this test is offered in the hope that schools will be stimulated to introduce periodic oral testing into their language programs, and to begin the establishment of norms for the semester-by-semester progress of their students in attaining an oral comand of French.

UNIVERSITY OF CALIFORNIA, LOS ANGELES

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY FRENCH REVIEW

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE U.S. OFFICE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM REQUIRES PERMISSION OF THE COPYRIGHT OWNER."